Anger Detection in Call Center Dialogues

Dimitris Pappas
Institute for Language and Speech Processing,
Epidavrou & Artemidos 6,
151 25 Maroussi, Greece
pappasd@aueb.gr

Ion Androutsopoulos
Department of Informatics,
Athens University of Economics and Business,
Patission 76, 104 34 Athens, Greece
ion@aueb.gr

Haris Papageorgiou
Institute for Language and Speech Processing,
Epidavrou & Artemidos 6,
151 25 Maroussi, Greece
xaris@ilsp.gr

***Abstract*—We present a method to classify fixed-duration windows of speech as expressing anger or not, which does not require speech recognition, utterance segmentation, or separating the utterances of different speakers and can, thus, be easily applied to real-world recordings. We also introduce the task of ranking a set of spoken dialogues by decreasing percentage of anger duration, as a step towards helping call center supervisors and analysts identify conversations requiring further action. Our work is among the very few attempts to detect emotions in spontaneous human-human dialogues recorded in call centers, as opposed to acted studio recordings or human-machine dialogues. We show that despite the non-perfect performance (approx. 70% accuracy) of the window-level classifier, its decisions help produce a ranking of entire conversations by decreasing percentage of anger duration that is clearly better than a random ranking, which represents the case where supervisors and analysts randomly select conversations to inspect.**

Keywords: anger detection, call centers, speech, machine learning

I. Introduction

Sentiment analysis [15], [16] is a flourishing research area, with applications, for example, in customer management systems, market movement forecasting, health and psychological care [18], [19]. Most sentiment analysis research, however, considers *texts*, for example social media posts [17] or product reviews [20]. Sentiment analysis for *spoken* utterances or dialogues has received less attention. Call centers, in particular, could benefit from sentiment analysis for speech. Thousands of phone calls are handled on a daily basis by large call centers, which struggle to ensure compliance with protocols and legislation that govern customer-agent interactions. Conversations where customers or, even worse, agents express excessive anger are one of the kinds of dialogues that need to be monitored, for example to determine when follow-up calls need to be made to customers, or advice needs to be offered to agents, marketing, or public relations departments.

In this paper, we report work towards automatically ranking call center conversations by the extent of anger they contain. We segment each conversation into contiguous non-overlapping windows, one second long each, and we train a Logistic Regression (LR) classifier [21] to classify each window as expressing anger or not. Once the classifier has been trained, we use it to rank any new set of conversations by decreasing percentage of windows that express (according to the LR classifier) anger. This way call center supervisors and analysts can focus on highly ranked conversations that may be more likely to require action. Our work is a first step towards a more general system intended to detect more emotions (e.g., surprise, joy, fear) [22] in call center dialogues, possibly in terms of real-valued dimensions (e.g., valence, arousal) [23].

Our LR classifier uses only features directly extracted from the speech signal (e.g., energy, pitch, MFCC features), unlike other approaches, discussed below, that also employ Automatic Speech Recognition (ASR) [24] to obtain transcriptions (texts) of the spoken utterances. Although transcriptions can be helpful in anger detection (e.g., when profanity is involved), ASR often performs poorly in phone conversations, especially emotionally intense conversations, and accurate speech recognizers are difficult to obtain in less widely spoken languages (our experiments were in Greek). Hence, there is value in investigating how accurately anger can be detected directly from speech. Furthermore, our approach does not require methods to separate voiced from unvoiced (e.g., silence, noise) speech segments, or methods to separate the voices of different speakers (customer and agent, in our case), which introduce their own errors. Most importantly, we experimented with real-world recordings between customers and human agents, obtained from the call center of a telephone provider, unlike work that uses acted speech recorded in studios [25]–[27], or recordings where customers interact with Interactive Voice Response (IVR) or Voice User Interface (VUI) systems [28]. We show experimentally that on a balanced development dataset (equal number of windows with and without anger), our LR classifier reaches an accuracy score of approximately 70%. We also show that despite its non-perfect accuracy, employing the LR classifier to rank unseen sets of conversations by the percentage of anger windows they contain (without balancing the numbers of windows) clearly outperforms a random ranker.

Overall, our main contributions are: (i) we present a method to classify fixed-duration windows of speech as expressing anger or not, which does not require ASR, utterance segmentation, or separating the utterances of different speakers and can, thus, be easily applied to real-world recordings; (ii) we introduce the task of ranking a set of spoken dialogues by decreasing percentage of anger duration, as a first step towards helping call center supervisors and analysts identify conversations that require further action; (iii) our work is among the very few attempts to detect emotions in spontaneous human-human dialogues recorded in call centers, as opposed to acted studio recordings or dialogues with IVR and VUI systems; (iv) we show that despite the non-perfect performance (approx. 70% accuracy) of the window-level classifier, its decisions help produce a ranking of entire conversations by decreasing

percentage of anger duration that is clearly better than a random ranking (which represents the case where supervisors and analysts randomly select conversations to inspect).

Section II below presents the datasets we used. Section III discusses the classifier that identifies windows expressing anger, also reporting experiments we performed to evaluate its performance. Section IV explains how the classifier is used to rank conversations by decreasing percentage of anger duration; it also presents experiments performed to evaluate the resulting rankings. Section V discusses related work. Section VI concludes and proposed directions for future work.

II. Data

The dataset we used contains 137 recorded customer-agent conversations, 9 hours and 30 minutes long in total, from the call center of a telephone provider company. All the conversations are in Greek. They can be divided into 7 kinds: *Mobile* (15 conversations, approx. 52 minutes in total), where the agent tries to promote mobile telephony; *Upgrade* (23 conversations, approx. 101 minutes), where the agent tries to convince a customer to upgrade to a new type of contract or service; *Telesales* (24 conversations, approx. 121 minutes), where the agent tries to attract a new customer; *Welcome* (17 conversations, approx. 64 minutes), where the agent welcomes a new customer; *Churns* (13 conversations, approx. 64 minutes), where customers ask to terminate their contracts; and *Information* (45 conversations, approx. 176 minutes), where customers ask for information. There are 76 calls where both the agent and the customer are female, 42 where the agent is female and the customer male, 12 where the agent is male and the customer female, and 7 calls where they are both male. All the conversations were recorded in one channel, with a sample rate of 8 kHz and 16 bits depth.

The first author listened to the conversations and annotated the speech segments where anger was expressed, using the ELAN annotation tool [10]. A weakness of our work is that we have not measured inter-annotator agreement. The first author, however, annotated the conversations twice, the second time correcting the annotations of the first pass, having also obtained a more consistent view of what counts as anger (having listened to all the conversations). Only 36 of the 137 conversations contained anger segments (Fig. 1, left).

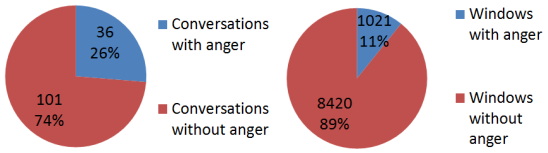

Fig. 1. Left: conversations with or without (any) manually annotated anger segments. Right: windows expressing or not expressing anger, according to the human annotation (for $t = 50\%$), in conversations with at least one manually annotated anger segment.

III. Window-Level Anger Detection

We segmented all the conversations into contiguous non-overlapping windows, one second long each (Fig. 2), in order to train the LR classifier to predict when a window expresses anger or not. The true (gold) category of each window was taken to be *Angry* if at least t percent of its duration overlapped with manually annotated anger segments (Fig. 3), and *Not Angry* otherwise. In most of our experiments, $t = 50\%$, though we also report experiments where different values of t were investigated. Even in the 36 conversations that contain at least one manually annotated anger segment, only 11% of the windows belong in the *Angry* category (Fig. 1, right), when $t = 50\%$. Hence, there is a severe category imbalance.

To address it, in the preliminary experiment of this section we under-sampled the majority category, i.e., we used all the windows of the Angry category and an equal number of randomly selected windows from the Not Angry category, during both training and testing.[1]. In the more realistic scenario of the next section, we also experiment without under-sampling the test data. We used the LibLinear [12] implementation of the LR classifier, with default settings.

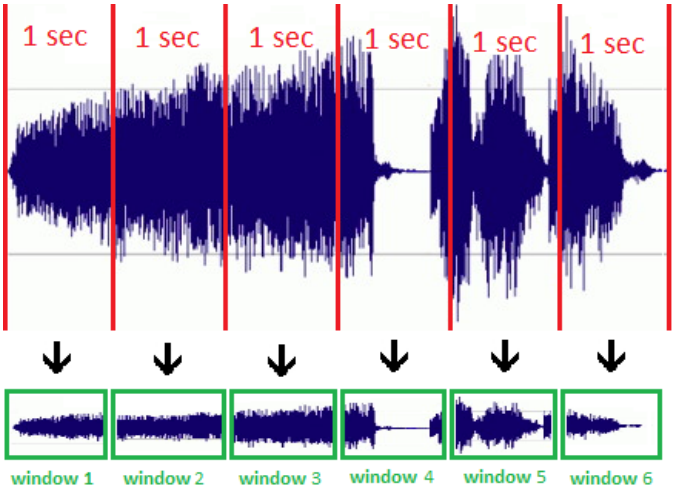

Fig. 2. Segmenting the waveform of a conversation into contiguous non-overlapping windows.

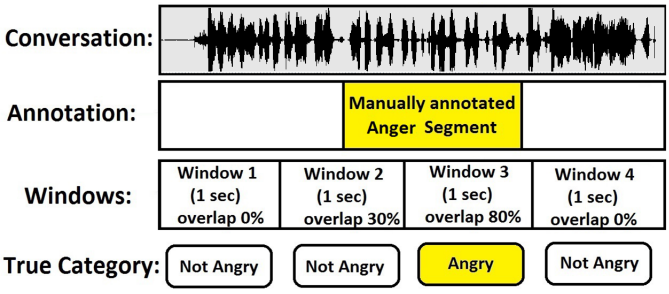

Fig. 3. A manually annotated anger segment, and contiguous non-overlapping windows with their true (gold) categories (bottom, for $t = 50\%$).

We used the OpenSmile toolkit [11] to construct a feature vector representation of each window. OpenSmile partitions the input speech segment (in our case, each window) into smaller *frames* (Fig. 4), extracts features (e.g., energy, pitch, MFCC features) from each frame, and then extracts additional aggregate ('functional') features (e.g., mean, maximum, minimum, standard deviation) over the frame features of the window (e.g., mean energy of the frames). The feature vector

[1]For each *Angry* window we kept a *Not Angry* window from the same conversation.

respresentation of each window contains the features of its frames (e.g., energy of first frame, energy of second frame etc.) and the aggregate features. The exact frame and aggregate features to be extracted are specified in a configuration file of OpenSmile, which also specifies the length and overlap of the frames. We used the configuration file (hence, exactly the same frame and aggregate features) of the INTERSPEECH 2009 Emotion Challenge [13] that leads to 384 features per window, using 25 msec frames, with a 60% overlap between consecutive frames. In previous work [14], we also experimented with alternative configuration files and different numbers of features, concluding that the 384 features were in practice the best.

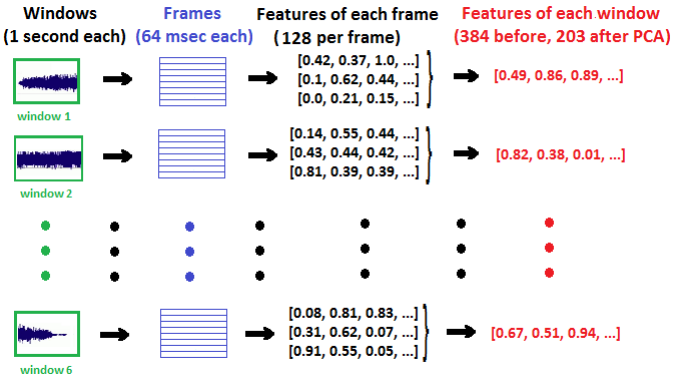

Fig. 4. Segmenting windows into frames, extracting features from each frame and (aggregate) features from all the frames of a window, and constructing a feature vector representation of each window.

We normalized the values of each feature to $[0, 1]$ using Min-Max scaling, and we then employed Principal Component Analysis (PCA) [9] to reduce the dimensionality of the feature space and obtain linearly independent features. We kept the 203 new features that corresponded to the eigenvectors with the highest eigenvalues; the sum of the eigenvalues of the 203 new features was 99% of the sum of all the eigenvalues.

Figure 5 shows the accuracy of the LR classifier on unseen test windows (correctly classified test windows divided by the total number of test windows), and the accuracy on the training windows the classifier has been trained on (training windows that have been used and also classified correctly, divided by the total number of training windows that have been used), as a function of the number of training windows used, with and without PCA (203 and 384 features, respectively). This experiment used the balanced form of our windows dataset (1021 *Angry* and 1021 *Not Angry* windows), with a leave-one-agent-out cross-validation. By the latter we mean that the experiment was repeated 21 times, as many iterations as the agents of our balanced dataset.[2] In each iteration, we used the windows from the conversations that involved a particular different agent as test windows, and the windows from all the other conversations as training windows; this ensures that the LR classifier does not overfit particular agents [3]. The accuracy scores are then (micro-) averaged over the 21 iterations.

The training accuracy (green curves) of Fig. 5 can be seen

[2]The entire dataset involves 56 agents, but only 21 of them participated in conversations with at least one *Angry* window.

[3]This was not a concern for customers, because each customer of the dataset participates in only one (training or test) conversation.

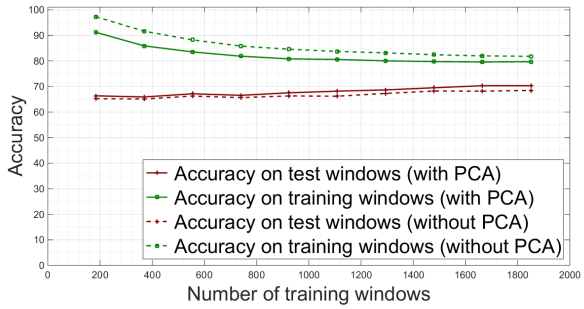

Fig. 5. Accuracy of the LR classifier on test and training windows, using a leave-one-agent-out cross-validation, with a balanced dataset (equal number of *Angry* and *Not Angry* windows), with and without PCA.

as an upper bound of the test accuracy (red curves), since classifiers usually perform better when classifying the instances they have been trained on, and worse when classifying unseen (test) instances. The training accuracy declines as more training windows are used, because it becomes more difficult for the classifier to overfit the training data. As more training windows are used, the test accuracy reaches 70.31% (with PCA), for 1,800 training windows. A further small improvement may be possible with more training windows, but test accuracy is unlikely to exceed 80% (with training accuracy viewed as an upper bound). PCA slightly reduces overfitting (the gap between training and test accuracy).

IV. Dialogue-Level Anger Detection

Large call centers handle thousands of customer-agent conversations on a daily basis. As already discussed, the conversations need to be monitored, for example to detect cases where follow-up calls need to be made (e.g., to customers that were particularly unhappy and threatened to break their contracts), or advice needs to be offered to agents (e.g., agents who frequently lose their temper), marketing, or public relations departments (e.g., when large numbers of customers express discontent during an advertising campaign). Because of the sheer number of the conversations, call center supervisors and analysts often resort to monitoring random samples of the conversations, perhaps involving particular age groups, customers who called several times etc. Tools that would help them identify conversations that require action would be particularly helpful, and conversations with excessive anger often belong in this category. Hence, in this part of our work we aimed to build a system that would allow a set of conversations (e.g., from a particular day) to be ranked by decreasing *anger charge*, which we define as the number of truly *Angry* windows divided by the total number of windows of a conversation. We hypothesize that conversations with high anger charge are more likely to require action. We do not actually test the latter hypothesis in this work, but we investigate the extent to which the LR classifier of the previous section can be used to correctly rank conversations by decreasing anger charge. We show experimentally that despite the non-perfect accuracy of the LR classifier (Fig. 5), its decisions can indeed help us produce a ranking (of entire conversations) that is cleary better than a random ranking (which represents the situation where supervisors and analysts select conversations randomly).

For the purposes of this part of our work, we used the 36 conversations that contained manually annotated anger segments (Fig. 1) and 36 randomly selected conversations that did not contain any anger segments. We then split the 72 conversations into training and test conversations, using an approximate 70%-30% split (53 training conversations, 19 test conversations), ensuring at the same time that no agent participated in both training and test conversations. The training conversations contained only 591 *Angry* windows and 12,658 *Not Angry* windows; hence, we again undersampled the *Not Angry* training windows to obtain an equal number of training windows from both categories. The test conversations contained 421 *Angry* windows and 4,489 *Not Angry* windows, but we *did not undersample* the *Not Angry* test windows to make the experiment more realistic.

Having trained the LR classifier on the windows from the training conversations, we used it to classify the windows of the test conversations as *Angry* and *Not Angry*, and then ranked the test conversations by decreasing anger charge, trusting the decisions of the classifier; we call this the *predicted ranking*. We also ranked the test conversations by decreasing anger charge, this time using the true (gold) categories of their windows (Fig. 3); we call this the *correct ranking*. Among the 19 test conversations, 13 did not contain any truly *Angry* windows, i.e., their true anger charge was zero, and 6 contained at least one *Angry* window.[4] Assuming that $A_1, A_2, \ldots, A_6$ are the labels of the 6 conversations (with at least one *Angry* window each) by decreasing true anger charge, and $N_1, N_2, \ldots, N_{13}$ are the labels of the other 13 conversations (with zero anger charge), the correct ranking (shown as a sequence of labels) is the following, where there are 13 Ns, and each N can be any (different) of $N_1, N_2, \ldots, N_{13}$, since the relative ordering of the conversations with zero true anger charge does not matter.

$$\text{correct ranking: } \langle A_1, A_2, A_3, A_4, A_5, A_6, N, N, \ldots, N \rangle$$

By contrast, when trusting the decisions of the LR classifier, the computed anger charge scores may be different, leading to a predicted ranking like the following:

$$\text{predicted ranking: } \langle A_2, N, A_1, \ldots, N, A_3, A_6, N, N \rangle$$

We use the Kendall rank correlation coefficient (Kendall's τ) [34] to compute how close a predicted ranking is to the correct one. The coefficient ranges from -1 (completely reverse rankings) to 1 (identical rankings). Intuitively, it considers how many pairwise swaps must be made to the elements of one of the rankings to become the same as the other ranking. Figure 6 shows Kendall's τ between the predicted and correct rankings of the 22 test conversations, as a function of t (minimum overlap of a window with a manually annotated anger segment, for the true category of the window to be *Angry*, as in Fig. 4). Also shown is the average Kendall's τ between 100 randomly created ranked lists and the correct ranking, again as a function of t. Clearly the automatically predicted rankings are better than the random ones. Also, the exact value of t does not affect much the quality of the predicted ranking, with Kendall's τ being approximately equal to 0.8 for all t values; the best τ score ($\tau = 0.86$) was obtained for $t = 0.2$.

[4]By contrast, among the 53 training conversations, 23 did not contain any *Angry* windows and 30 contained at least one *Angry* window.

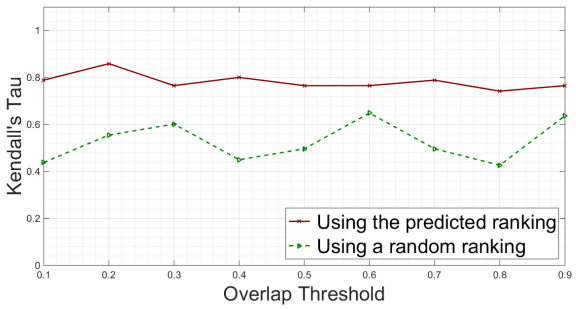

Fig. 6. Kendall's τ between the predicted and correct rankings of conversations by decreasing anger charge, as a function of t (min. overlap of a window with a manually annotated anger segment, for the true category of the window to be *Angry*). Also shown is the average Kendall's τ between 100 randomly created ranked lists and the correct ranking, as a function of t.

To obtain a clearer view of the performance of our approach, we also considered a simplified scenario, where the goal is to rank all the conversations that contain at least one *Angry* window (without considering their particular anger charge scores) above all the conversations that do not contain any *Angry* windows. In this case, the correct ranking for our 19 test conversations is the following, where there are 6 As and 13 Ns, each A can be any (different) of $A_1, A_2, \ldots, A_6$, since the relative ordering of the conversations with non-zero anger charge does not matter, and similarly for the Ns.

$$\text{simplified correct ranking: } \langle A, \ldots, A, N, N, \ldots, N \rangle$$

A predicted ranking may be the following, where again we use only two labels (A, N):

$$\text{simplified predicted ranking: } \langle A, N, A, \ldots, N, A, A, N, N \rangle$$

In the simplified scanario, we can plot the Average Interpolated Precision (AIP) [29] at different recall levels, as in Information Retrieval tasks.[5] Figure 7 shows the resulting AIP scores of the predicted rankings and the average AIP scores of 100 randomly created ranked lists. Our method is clearly better than a random ranker by a wide margin.

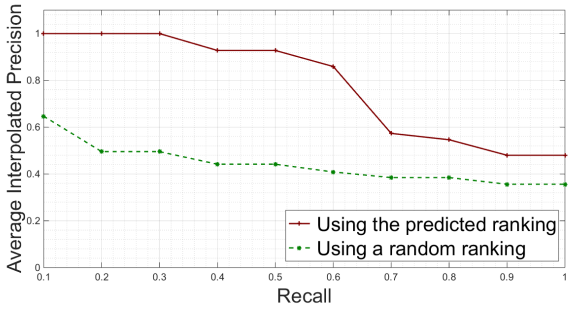

Fig. 7. Average Interpolated Precision (AIP) at different recall levels, for the simplified scenario, using the predicted and random rankings of conversations.

V. Related Work

Burkhardt et al. [1] used recordings from a German voice portal where customers reported problems with their phone

[5]In our case, each A corresponds to a relevant retrieved document of an Information Retrieval task, and each N to an irrelevant document.

connections; it is unclear to us if the customers interacted with automatic IVR systems only, or both IVR systems and human agents. Burkhardt et al. extracted acoustic features (namely pitch, intensity, duration, and prosodic features) aiming to detect angry utterances. Using Support Vector Machines (SVMs), they managed to achieve an F1 score of 70%. Unlike our work, they did not classify windows of fixed duration, but entire dialogue turns, which requires methods to identify dialogue turns in the speech signal. They also discarded 'garbage' dialogue turns not directed to the voice portal, but it is unclear how these turns could be automatically detected in practice. Furthermore, unlike our two categories (*Angry*, *Not Angry*), they used five levels (categories) of anger intensity, which makes manual annotation more difficult.

Gupta et al. [2] used two pipelines to detect angry, happy, and neutral utterances. The first pipeline used features extracted directly from the speech signal, while the second one used ASR to extract features from text. Gupta et al. tested the first pipeline on the LDC Emotional Speech Database [6], and both pipelines on real-word recordings from call centers. The LDC Database contains acted dialogues and, hence, is not directly relevant to our work. On the call center recordings, both pipelines achieved (separately) an accuracy of 60% when detecting angry utterances, whereas the corresponding accuracy of a system that combined (with weighted voting) both pipelines was 80%; hence, adding ASR is indeed beneficial. The 60% accuracy of the first pipeline (which uses GMMs [30]) is lower than the 70.31% of our LR classifier (which also uses only acoustic features), but the results are not comparable, because we use different datasets and Gupta et al. classify utterances (requiring an utterance segmentation method), whereas we classify windows of fixed duration.

Vidrascu et al. [3] used spontaneous dialogues, in French, recorded in a call center. They classified utterances (again requiring utterance segmentation), also assuming that the utterances of the agents and clients are separate (e.g., recorded in separate channels). Using SVMs they achieved 75% accuracy in distinguishing angry from neutral utterances of the agent, and 80.2% accuracy in distinguishing angry from neutral utterances of the client. By contrast, our approach does not require utterance segmentation, nor separating the utterances of different speakers. Furthermore, the recordings of Vidrascu et al. involve only 7 agents (for 688 dialogues), compared to the 56 agents of our recordings (for 137 dialogues); hence, there is smaller risk of overfitting particular agents in our work.

In other work, Vidrascu et al. [4] used a 10-hour dialogue corpus recorded in a French Medical emergency call center. They achieved 82% accuracy in distinguishing dialogue turns expressing negative emotions (e.g., fear, anger, sadness) from turns expressing positive emotions (e.g., interest, relief, compassion), using acoustic and lexical (via ASR) features, several feature selection techniques (e.g., OneR, gain ratio), SVMs or Logistic Model Trees [31]. Again, the recordings involve only 6 agents (for 404 dialogues), and dialogue turn segmentation methods are required.

Lee et al. [5] used recordings of human-*machine* dialogues from a commercial call center application, aiming to distinguish utterances with negative emotions from utterances with non-negative emotions. They employed pitch and energy features from the acoustic signal, feature selection, dimensionality reduction via PCA, and a Linear Discriminant Classifier [32] or, alternatively, a k-nearest neighbour classifier. They achieved 80% and 75.8% accuracy on male and female utterances, respectively, using separate classifiers for male and female speakers, having manually separated male from female utterances. Again, utterance segmentation is required, as well as methods to distinguish male from female utterances.

Pohjalainen et al. [7] experimented with the widely used Berlin Database (also known as Emo-DB), which contains *acted* stand-alone German utterances, aiming to distinguish angry from non-angry utterances. They added three types of noise (car, factory, babble), to make the dataset more realistic. Using GMMs they achieved 93% accuracy. The dataset, however, is far from spontaneous dialogues recorded in call centers.

Polzehl et al. [8] used customer utterances from dialogues with an English and a German IVR, and children utterances from a German Wizard of Oz experiment with a robotic pet. They employed acoustic features (pitch, loudness, MFCC, spectral information, formants, intensity), ASR, feature selection (gain ratio), and an SVM (with an RBF kernel) or, alternatively, a multilayer perceptron [33]. All the recordings were already segmented in utterances. The best scores of Polzehl et al. were 79.0% accuracy for the German IVR, 78.2% for the English IVR, and 75.3% for the German Wizard of Oz.

VI. Conclusions and Future Work

We presented a method to classify fixed-duration windows of speech as expressing anger or not, which does not require ASR, utterance segmentation, or separating the utterances of different speakers and can, thus, be easily applied to real-world recordings. We also introduced the task of ranking a set of spoken dialogues by decreasing anger charge (percentage of *Angry* windows), as a step towards helping call center supervisors and analysts identify conversations requiring further action. Our work is among the very few attempts to detect emotions in spontaneous human-human dialogues recorded in call centers, as opposed to acted studio recordings or human-machine dialogues. We showed that despite the non-perfect performance (approx. 70% accuracy) of the window-level classifier, its decisions help produce a ranking of entire conversations by decreasing anger charge that is clearly better (in terms of Kendall's τ with the correct ranking) than a random ranking, which represents the case where supervisors and analysts randomly select conversations to inspect. We also considered a simplified scenario, where the goal was to rank all the conversations that contained at least one *Angry* window (without considering their anger charges) above all the conversations that did not contain any *Angry* windows (zero anger charge), with results (in terms of Average Interpolated Precision) indicating that our method is again clearly better than a random ranker, by a wide margin.

An obvious extension of our work would be to consider additional emotions (e.g., surpise, joy, fear) in call center dialogues, possibly in terms of real-valued dimensions (e.g., valence, arousal). It would also be worth measuring inter-annotator agreement and experimenting with larger datasets, though manual annotation (especially with multiple annotators) is costly and call center recordings are often difficult to obtain for research purposes due to privacy concerns. Another extension would be to incorporate ASR (for languages

where reliable speech recognizers are available), since previous research indicates that ASR can lead to improved emotion detection. Noise cancellation could also be added to reduce background noise, and different recording channels could be used for the clients and agents, since different actions may be desirable when clients or agents express excessive anger.

REFERENCES

[1] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, R. Huber, "Detecting real life anger" Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing pp. 47614764, Taipei, Taiwan, 2009.

[2] P. Gupta, N. Rajput, "Two-stream emotion recognition for call center monitoring.", Proceedings of the Annual Conference of the International Speech Communication Association INTERSPEECH 2007, Brisbane, Australia, 2007.

[3] L. Vidrascu, L. Devillers, "Real-life emotion representation and detection in call centers data", Affective computing and intelligent interaction, 3784:739-746, Springer, 2005.

[4] L. Vidrascu, L. Devillers, "Detection of real-life emotions in call centers.", Proceedings of the Annual Conference of the International Speech Communication Association INTERSPEECH 2005, Pittsburgh, Pennsylvania, 10:1841–1844, 2005.

[5] C. M. Lee, S. Narayanan, R. Pieraccini, "Recognition of negative emotions from the speech signal", IEEE Workshop on Automatic Speech Recognition and Understanding ASRU 01, Madonna di Campiglio, Italy, 2001.

[6] M. Liberman, K. Davis, M. Grossman, N. Martey, J. Bell, "Emotional prosody speech and transcripts", Linguistic Data Consortium, Philadelphia, 2002.

[7] J. Pohjalainen, P. Alku, "Automatic detection of anger in telephone speech with robust autoregressive modulation filtering.", Acoustics, Speech and Signal Processing ICASSP 2013, pp. 7537-7541, 2013.

[8] T. Polzehl, A. Schmitt, F. Metze, M. Wagner, "Anger recognition in speech using acoustic and linguistic cues", Speech Communication, 53(9-10):11981209, 2011.

[9] I. T. Jolliffe, "Principal Component Analysis, Second Edition", Encyclopedia of Statistics in Behavioral Science, 30(3):487, 2002.

[10] H. Sloetjes, P. Wittenburg, "Annotation by category - ELAN and ISO DCR", Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech,Morocco, pp. 816820, 2008.

[11] F. Eyben, F. Weninger, F. Gross, B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor", Proceedings of the 21st ACM international conference on Multimedia MM'13, Barcelona, Catalunya, Spain, pp. 835838, 2013.

[12] R.Fan, K. Chang, C. Hsieh, "LIBLINEAR: A library for large linear classification", The Journal of Machine Learning, 9:18711874, 2008.

[13] Bj. Schuller, St. Steidl, A. Batliner, "The INTERSPEECH 2009 emotion challenge.", Proceedings of the Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, pp. 312-315, 2009.

[14] D. Pappas, "Emotion recognition in spoken dialogues", MSc thesis, Department of Informatics, Athens University of Economics and Business, 2015.

[15] B. Liu, "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, 5(1):1–167, Morgan & Claypool Publishers, 2012.

[16] B. Pang, l. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, 2(1-2):1–135, Now Publishers Inc., 2008.

[17] S. Rosenthal, Pr. Nakov, Sv. Kiritchenko, S.M. Mohammad, A. Ritter, V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter.", Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, Denver, Colorado, 2015.

[18] C. Oh, O.R.L. Sheng, "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement", Proceedings of the International Conference on Information Systems (ICIS), pp. 119, 2011.

[19] J.P. Pestian, P. Matykiewicz, M. Linn-Gust, Br. South, Oz. Uzuner, J. Wiebe, K.B. Cohen, J. Hurdle, Chr. Brew, "Sentiment analysis of suicide notes: A shared task.", Biomedical informatics insights, 5(Suppl 1):3, NIH Public Access, 2012.

[20] M. Pontiki, D. Galanis, H. Papageogiou, S. Manandhar, I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis.", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, 2015.

[21] DW. W. Hosmer, S. Lemeshow, Applied logistic regression, John Wiley & Sons, 2000.

[22] P. Ekman, "An argument for basic emotions", Cognition & emotion, 6(3):169-200, 1992.

[23] D. C. Rubin, J. M. Talarico, "A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words", Memory (Hove, England), 17(8):802808, 2009.

[24] X. Huang, A. Acero, H. Hon, R. Foreword By-Reddy, "Spoken language processing: A guide to theory, algorithm, and system development", Prentice Hall PTR, pp. 933, 2001.

[25] F. Burkhardt, M. van Ballegooy, R. Englert, R. Huber, "An emotion-aware voice portal.", In Electronic Speech Signal Processing Conference ESSP, pp. 123-131, Prague, Czech Republic, 2005.

[26] B. Schuller, G. Rigoll, M. Lang, "Hidden markov model based speech emotion recognition", International Conference on Acoustics, Speech, and Signal Processing ICASSP 2003, (2):14, Hong Kong, China, 2003.

[27] D. Ververidis, C. Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm", IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005.

[28] C. M. Lee, S. Narayanan, R. Pieraccini, "Recognition of negative emotions from the speech signal", IEEE Workshop on Automatic Speech Recognition and Understanding ASRU 01, Madonna di Campiglio, Italy, 2001.

[29] C. D. Manning, P. Raghavan, H. Schtze, "Introduction to Information Retrieval", Cambridge University Press, Vol. 1, 2008.

[30] C. M. Bishop, "Pattern recognition and machine learning", Pattern Recognition, (Vol. 4), Springer, 2006.

[31] N. Landwehr, M. Hall, E. Frank, "Logistic model trees", Machine Learning, 59(1-2):161205, 2005.

[32] S. Chen, X. Yang, "Alternative linear discriminant classifier", Pattern Recognition, 37(7):1545-1547, 2004.

[33] S. Haykin, "Neural Networks and Learning Machines", Pearson Prentice Hall New Jersey USA 936 pLinks, 3(10):906, 2008.

[34] M. G. Kendall, "Rank Correlation Methods", Charles Griffin & Company Limited, 1948.